DNA 甲基化微阵列的非参数贝叶斯聚类算法

张 林 [1]　刘 辉 [1]

摘　要　面向 Illumina GoldenGate 甲基化微阵列数据提出了一种基于模型的聚类算法. 算法通过建立贝塔无限混合模型, 采用 Dirichlet 过程作为先验, 实现了基于数据和模型的聚类结构的建立, 实验结果表明该算法能够有效估计出聚类类别个数、每个聚类类别的混合权重、每个聚类类别的特征等信息, 达到比较理想的聚类效果.

关键词　DNA 甲基化微阵列, Dirichlet 过程, 贝塔混合模型, 吉布斯抽样

引用格式　张林, 刘辉. DNA 甲基化微阵列的非参数贝叶斯聚类算法. 自动化学报, 2012, **38**(10): 1709−1713

DOI　10.3724/SP.J.1004.2012.01709

Nonparametric Bayesian Clustering Methods of DNA Methylation Microarray

ZHANG Lin[1]　LIU Hui[1]

Abstract　A model based clustering algorithm for Illumina GoldenGate microarray data is proposed in this paper. By infinite beta mixture model and by adopting Dirichlet process as prior knowledge, the cluster structure can be defined based on model and data. Simulation results demonstrate that this methodology can estimate the number of clusters, the cluster mixing weight and the own characteristic of each cluster, and can reach relatively ideal clustering effect.

Key words　DNA methylation microarray, Dirichlet process, beta mixture model, Gibbs sampling

Citation　Zhang Lin, Liu Hui. Nonparametric Bayesian clustering methods of DNA methylation microarray. *Acta Automatica Sinica*, 2012, **38**(10): 1709−1713

人类基因组中, 70 % ∼ 80 % 的 CpG 双核苷酸处于 DNA 甲基化状态. DNA 甲基化在基因表达调控中发挥着重大作用. 非甲基化的 CpG 并非均匀分布, 而是呈现出局部聚集的倾向, 从而形成了一些 GC 含量较高、CpG 双核苷酸相对聚集的区域, 通常称为 CpG 岛. 根据 Gardiner-Garden 等的定义, CpG 岛被定义为一段长度不小于 200 bp、GC 含量不低于 50 %、CpG 含量与期望含量之比值不低于 0.6 的区域[1]. 随着研究的不断深入, 人们发现甲基化的 CpG 存在于印迹基因、失活的 X 染色体甚至正常的体细胞中, 而某些 CpG 岛的异常甲基化伴随着癌症等疾病的发生[2−3]. 近年来, 随着高通量的 DNA 甲基化检测技术的出现, DNA 甲基化的数据挖掘技术得到了很大的发展. 一系列预测工具一方面成为实验检测技术的有力补充, 另一方面也反映了 DNA 甲基化本身有规律可寻, 因而启发了研究者们对 DNA 甲基化内在机制的探索. 例如, Ang 等根据 CpG 岛甲基化表型考虑将结直肠癌分为三种亚型, 通过对 28 例正常的结肠黏膜和 91 例结肠癌样本的 Illumina GoldenGate 甲基化阵列进行系统聚类分析发现, 不同的结直肠癌亚型存在着不同的临床病理学特征[4].

DNA 甲基化微阵列表达数据通常可表示为矩阵, 矩阵中各行表示甲基化位点, 各列表示样本. 聚类, 即无监督分类, 作为一种将数据集划分为若干组或类的过程, 已被广泛应用于数据挖掘、统计学、机器学习、生物信息学等多个领域. 通过聚类, 同一组内的数据对象具有较高的相似度, 而不同组中的数据对象则不相似. 相似与不相似基于数据对象描述属性的取值确定. 对于 DNA 甲基化微阵列数据, 聚类算法可以在行和列两个方向实现: 对行 (甲基化位点) 实现聚类, 可以发现差异甲基化位点或功能相关位点并得出生物学方面的结论; 对列 (样本) 实现聚类, 主要用于诊断疾病 (如癌症) 的亚类或发现新的疾病种类. 相关的聚类方法主要有划分类方法、分层类方法、基于密度类方法、基于网格类方法和基于模型类方法. 文献 [5] 指出基于模型的聚类算法更适用于分析 DNA 甲基化数据.

目前很多算法不能根据数据的真实模型自动确定聚类数目, 文献 [6] 提出的 RPMM (Recursively partitioned mixture model) 通过建立贝塔混合模型描述 Illumina 甲基化微阵列表达谱, 并通过改进的 BIC (Bayesian information criterion) 准则确定数据中隐含的聚类数目. 该算法的实质为: 首先假定聚类数目可能的取值区间, 在此假设的基础上分别建立模型, 再通过改进的 BIC 准则对上述几个模型进行选择. 显而易见, 应用该方法进行聚类分析时必须预先估计聚类数目的可能值, 导致选择的最佳聚类数目将局限于事先选取的取值区间中, 因而存在陷入局部最佳的风险.

本文的目的是建立一种可直接基于数据和模型确定数据聚类结构的方法. 从发现聚类模型中的聚类类别个数入手, 通过建立 Dirichlet 过程贝塔混合模型 (Dirichlet process beta mixture model, DPBMM) 实现 DNA 甲基化微阵列数据的聚类分析. 鉴于 Dirichlet 过程混合模型中混合模型的个数无需预先确定而是采用 ∞ 的先验知识[7−9], 因而避免了常见的聚类分析过程中对聚类类别个数的人为干预.

1　Illumina GoldenGate 甲基化微阵列数据

Illumina 基于 Infinium 和 GoldenGate 两种平台, 相应推出了人类全基因组甲基化芯片和 GoldenGate 甲基化肿瘤芯片. 其中, Illumina GoldenGate 甲基化芯片可以自定义甲基化位点, 能够提供 700 多个肿瘤相关基因的 1 500 多个 CpG 岛. 经过双色荧光杂交之后, 可获得各甲基化位点甲基化表达水平和非甲基化表达水平. DNA 甲基化芯片中的甲基化水平值则代表了甲基化等位基因密度 (M) 与非甲基化等位基因密度 (U) 的比值, 常称作 *beta* 值, 如式 (1) 所示.

$$beta = \frac{\max(M,0)}{\max(M,0)+\max(U,0)+100} \tag{1}$$

可见, *beta* 是 $[0,1]$ 之间的连续值. 因此, 可建立贝塔分布的混合模型以描述 DNA 甲基化芯片中各等位基因的甲基化水平[10].

收稿日期 2011-08-22　录用日期 2012-03-28
Manuscript received August 22, 2011; accepted March 28, 2012
中央高校业务专项基金 (2010QNA50, 2010QNA47)
Supported by the Fundamental Research Funds for the Central Universities (2010QNA50, 2010QNA47)
本文责任编委 姚力
Recommended by Associate Editor YAO Li
1. 中国矿业大学信息与电气工程学院 徐州 221116
1. School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116

2 DNA 甲基化微阵列混合模型

2.1 DNA 甲基化微阵列贝塔混合模型

考虑基于 Illumina GoldenGate 甲基化微阵列数据展开的聚类问题, 假设 $X = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$ 表示 n 个样本的 DNA 甲基化微阵列表达水平. 第 i 个样本 $\boldsymbol{X}_i = \{x_{i1}, x_{i2}, \cdots, x_{iJ}\}$ 则包括了其 J 个甲基化位点上的表达水平, 这些值均分布于 $[0,1]$ 之间, 因此可用贝塔分布或者多个贝塔分布的混合模型对该类数据进行描述:

$$f(\theta_i) = \sum_{k=1}^{\infty} \pi_k \prod_{j=1}^{J} \frac{x_{ij}^{\alpha_{kj}-1}(1-x_{ij})^{\beta_{kj}-1}}{\mathrm{B}(\alpha_{kj}, \beta_{kj})} \tag{2}$$

其中, π 表示混合模型中的混合权重, k 表示混合模型中聚类的个数. $\theta = \{\alpha, \beta\}$ 则表示贝塔分布的参数, 选取不同的 θ 即可描述不同的贝塔分布, 如图 1 所示.

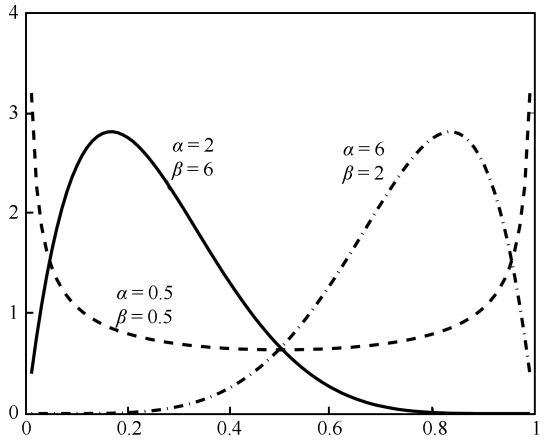

图 1 参数 θ 不同时贝塔分布的概率密度函数曲线

Fig. 1 Curves of beta probability density function based on different θ

2.2 Dirichlet 分布

所谓 Dirichlet 分布是多项式分布的共轭分布, 一般表示为多项式分布中的概率参数分布. 因此, Dirichlet 分布可以看作是分布上的分布, 其形式如下:

$$\mathrm{Dir}(\theta|\alpha, M) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{K}\Gamma(\alpha m_i)} \prod_{i=1}^{K} \theta_i^{\alpha m_i - 1} \tag{3}$$

其中, $\Gamma(\alpha)$ 为伽马函数, m_i 满足约束条件 $\sum_{i=1}^{K} m_i = 1$, 并且 $0 \le m_i \le 1$.

2.3 Dirichlet 过程

Dirichlet 过程可以简单地看作 Dirichlet 分布向连续空间的扩展, 因此 Dirichlet 过程的性质与 Dirichlet 分布十分相似. 一般, Dirichlet 过程可表示为 $G \sim DP(\tau, G_0)$, 其中 G_0 是基分布, τ 则是一个表示集中度的参数, 展示着 G 向 G_0 的逼近程度, 并有 $\mathrm{E}[G] = G_0$. 可见, 一个满足 Dirichlet 过程的分布是一个离散的分布, 但这种离散的分布却无法用有限个参数加以描述, 所以 Dirichlet 过程通常可理解为一种分布上的分布. 基于贝叶斯理论, 参数为 θ 的 Dirichlet 过程可表示为

$$\begin{aligned} X_i|\theta_i &\sim f(\theta_i) \\ \theta_i|G &\sim G \\ G|\tau, G_0 &\sim DP(\tau, G_0) \end{aligned} \tag{4}$$

基于 Dirichlet 过程, 模型 (2) 中的参数 k 可被赋予 ∞ 的先验知识, $f(\theta_i)$ 如式 (2) 所示.

2.4 后验分布计算

由于 Dirichlet 过程以概率 1 趋于离散化, 因此对于 ∞ 个参数 θ_i 而言, 其中必有一些参数相等. 假设 $\phi = \{\phi_1, \cdots, \phi_k\}$ 表示 θ_i 中互不相同的 k 组取值, 同时引入 $\boldsymbol{s} = \{s_1, s_2, \cdots, s_n\}$ 表示 n 个样本的聚类标签, 则 $\theta = \{\theta_i : i = 1, \cdots, m\}$ 可表示为 $\{\phi_1, \cdots, \phi_k, s_1, \cdots, s_m\}$. 采用 $n_i, i = 1, \cdots, k$ 来表示第 i 个聚类中包含的样本数目, 即 $s_j = i$ 的个数. 而 $n_{-ij}, i = 1, \cdots, k$ 则表示不含第 i 个样本的其他样本经过聚类后第 j 类别中包含的样本数目.

为保证式 (2) 中参数 $\alpha > 0$ 和 $\beta > 0$, 将其分别变换成 L_α 和 L_β, 如式 (5) 所示.

$$\alpha = \exp(|L_\alpha|), \quad \beta = \exp(|L_\beta|) \tag{5}$$

则 $G_0(\alpha, \beta) = \mathrm{N}(0, \sigma_\alpha^2)\mathrm{N}(0, \sigma_\beta^2)$, 其中 $\mathrm{N}(\mu, \sigma^2)$ 是期望值为 μ, 方差为 σ^2 的正态分布. π 则遵循如式 (6) 所示的 Dirichlet 分布作为先验分布.

$$\pi \sim \mathrm{Dir}\left(n_1 + \frac{\tau}{k}, \cdots, n_k + \frac{\tau}{k}\right) \tag{6}$$

Dirichlet 过程存在多种描述方法, 而 Escobar 等[11] 首次采用吉布斯采样基于 Blackwell-MacQueen 的 Polya urn 模型[12] 实现了后验分布密度估计. 本文中 $s_i|s_1, \cdots, s_{i-1}, i = 1, \cdots, n$ 的条件先验分布如下:

$$\begin{aligned} &P(s_1 = 1) = 1 \\ &\begin{cases} P(s_i = j|s_1, \cdots, s_{i-1}) = \dfrac{n_{-i,j}}{\tau + i - 1}, \ j = 1, \cdots, k_i \\ P(s_i = k_i + 1|s_1, \cdots, s_{i-1}) = \dfrac{\tau}{\tau + i - 1} \end{cases} \end{aligned} \tag{7}$$

$\theta_i|\theta_{-i}$ 的条件先验分布为

$$\begin{aligned} \theta_1 &\sim G_0(\theta_1) \\ \theta_i|\theta_1 \cdots \theta_{i-1} &\sim \frac{\tau}{\tau + i - 1} G_0(\theta_i) + \\ &\sum_{j=1}^{k} n_j \frac{1}{\tau + i - 1} \delta_{\Phi_j}(\theta_i), \quad i \ge 1 \end{aligned} \tag{8}$$

由此可得出 θ_i 的后验分布为

$$\begin{aligned} p(\theta_i|\theta_{-i}, \boldsymbol{s}_{-i}, X) &\propto q_{i,0} G_i(\theta_i) + \sum_{j=1, j\ne i}^{n-1} q_{i,j} \delta_{\theta_j}(\theta_i) = \\ &q_{i,0} G_i(\theta_i) + \sum_{j=1}^{k} n_{-i,j} q_{i,j} \delta_{\Phi_j}(\theta_i) \end{aligned} \tag{9}$$

2.5 算法实现

可见, 式 (4) 中 Dirichlet 过程的基分布 G_0 与式 (2) 并不满足共轭的性质, 因此式 (9) 中 $q_{i,0}$ 很难计算, 从分布 G 中采样同样难以实现. MacEachern 等设计了一种称为 "no-gaps" 的算法[13], 可避免从式 (9) 采样, 从而解决非共轭先验分布积分不可求的情况.

"no-gaps" 算法将参数 $\phi = \{\phi_1, \cdots, \phi_k\}$ 扩展为最大可能聚类类别数 n, 即样本个数.

$$\{\underbrace{\phi_1, \cdots, \phi_k}_{\phi_F}, \underbrace{\phi_{k+1}, \cdots, \phi_n}_{\phi_E}\} \tag{10}$$

其对于这样 n 个聚类类别, 存在 $j = \{1, \cdots, k\}$ 时 $n_j > 0$, $j = \{k+1, \cdots, n\}$ 时 $n_j = 0$. 因此, 对应于 $\phi = \{\phi_1, \cdots, \phi_k\}$ 是实际存在的类别的参数集, 也称 ϕ_F 为 "满聚类"; 而 $\phi = \{\phi_{k+1}, \cdots, \phi_n\}$ 则表示可能出现的类别的参数集, 也称 ϕ_E 为 "空聚类".

最后, 基于吉布斯抽样、Metropolis 的 MCMC (Markov chain Monte Carlo) 方法抽样实现上述模型. 第 t 次吉布斯抽样更新过程如下[14]:

1) 对每个样本 $i = \{1, \cdots, n\}$, 以 $\boldsymbol{s}^{t-1}$, k^{t-1}, ϕ^{t-1}, π^{t-1}, τ^{t-1} 为基础更新聚类标签 $\boldsymbol{s}$, 重复步骤 a) 和 b).

a) 如果 $n_{s_i} > 1$, 基于式 (7) 进行抽样, 此时 $k_{-i} = k$;

b) 如果 $n_{s_i} = 1$, 则以 $1 - 1/k$ 的概率保持 s_i 不变, 以 $1/k$ 的概率对 $\boldsymbol{s}$ 重新排列, 使 $s_i = k$, 再基于式 (7) 进行抽样, 但此时 $k_{-i} = k - 1$;

2) 以 $\boldsymbol{s}^t$, k^t, ϕ^{t-1}, π^{t-1}, τ^{t-1} 为基础更新聚类模型参数 ϕ.

对每个 "满聚类" $i = \{1, \cdots, k\}$, ϕ_i 的后验概率分布如下:

$$\begin{aligned} &p(\Phi_i|\Phi_{-i}, \boldsymbol{s}, X, \pi) \propto \\ &\quad p(X_{m:s_m=s_i}|\Phi, \boldsymbol{s}, \pi) p(\Phi_i|\Phi_{-i}, \boldsymbol{s}, \pi) = \\ &\quad p(\Phi_i) \prod_{m:s_m=s_i} \prod_{j=1}^{J} \frac{x_{mj}^{\alpha_{kj}-1}(1-x_{mj})^{\beta_{kj}-1}}{B(\alpha_{kj}, \beta_{kj})} = \\ &\quad G_0 \prod_{m:s_m=s_i} \prod_{j=1}^{J} \frac{x_{mj}^{\alpha_{kj}-1}(1-x_{mj})^{\beta_{kj}-1}}{B(\alpha_{kj}, \beta_{kj})} \end{aligned} \tag{11}$$

对每个 "空聚类" $i = \{k+1, \cdots, n\}$, ϕ_i 的先验分布及后验分布均为 G_0.

3) 以 $\boldsymbol{s}^t$, k^t, ϕ^t, π^{t-1}, τ^{t-1} 为基础更新混合权重 π. 参数 π 的后验概率分布如式 (6) 所示, 其中 $n_k = \sum_{i=1}^{N} \delta(s_i, k)$.

4) 以 $\boldsymbol{s}^t$, k^t,ϕ^t, π^t, τ^{t-1} 为基础更新 Dirichlet 过程的集中度参数 τ.

3 实验结果分析

3.1 实验数据

本文在 Matlab 2009b 环境中实现基于 Dirichlet 过程的混合聚类模型的算法. 为了验证 Dirichlet 过程混合聚类模型算法的有效性, 进行了两组实验. 在第 1 组实验中, 按照式 (2) 中的贝塔混合模型产生仿真数据. 仿真数据包括 200 个样本, 每个样本包括 15 个表达值. 200 个样本分属于 4 个聚类类别, 每个聚类类别的贝塔参数 L_α 和 L_β 分别由 15 维正态分布产生, 正态分布的期望值均为 0, 协方差分别为对角线上元素是 5 和 6 的 15×15 对角阵. 产生的仿真数据映射为色图, 如图 2 所示.

在第 2 组实验中, 采用了文献 [6] 中 Case I 的 Illumina GoldenGate DNA 甲基化数据来验证上述聚类算法的有效性. 该数据集中包括 100 个样本, 每个样本提供 1 413 个表达值属性, 这些值分布于 [0, 1] 之间. 这 100 个样本来自 5 个聚类类别, 每个聚类类别以 0.2 的概率出现.

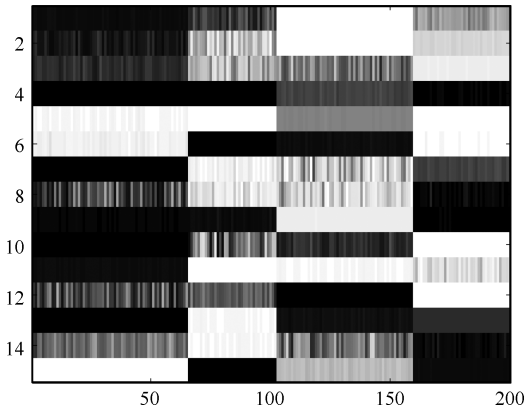

图 2　仿真数据映射为色图图谱

Fig. 2　Colormap plot of the simulated data set

3.2 实验结果

按照本文提出的 DPBMM 算法对上述两组数据集进行聚类分析. 其中聚类类别个数 k 采用 ∞ 作为先验知识, 但考虑到样本中样本数量有限, 因此聚类类别个数 k 的先验值设定为样本数量. 在第 1 组实验中, 经过 300 次 "burn-in" 吉布斯抽样, 聚类类别个数 k 保持在 4 不变, 如图 3 所示. 这与产生仿真数据模型的设定一致.

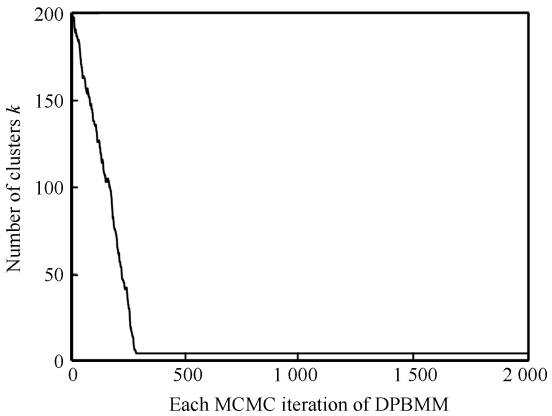

图 3　吉布斯抽样过程中聚类数目 k 的变化

Fig. 3　Number of clusters k in each Gibbs sampling iteration

对该数据集中的 200 个样本, 计算其查准率和查全率[15]发现, 经过 300 次 "burn-in" 吉布斯抽样之后, 查准率几乎接近 1, 如图 4 所示.

在第 2 组实验中, 首先按照文献 [6] 中的方法进行降维. 计算 100 个样本在每个甲基化位点上的基因表达水平的方差, 保留其中方差最大的 J 个甲基化位点参与聚类分析. 本文在 $J = 25$ 和 $J = 50$ 两种情况下分别开展聚类分析, 推断

出的聚类类别个数如表 1 所示. 可见, DPBMM 算法在推断聚类类别个数方面能够获得很好的效果.

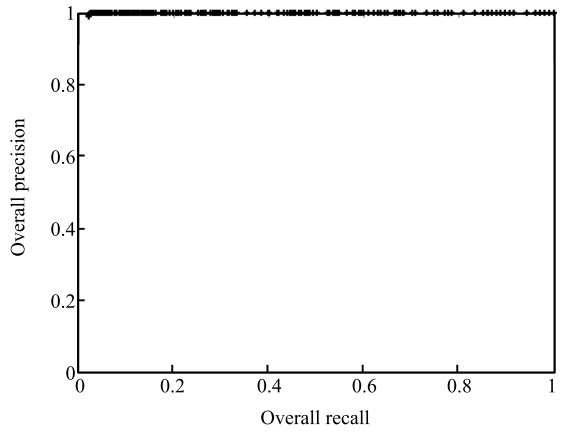

图 4 单次 DPBMM 迭代的整体查准率 – 整体查全率曲线

Fig. 4 The curve of overall precision – overall recall in one DPBMM iteration

表 1 基于 RPMM 和 DPBMM 得出的聚类数目统计

Table 1 Number of clusters obtained from RPMM and DPBMM in simulated data

方法	J	中值	期望	标准差
使用 BIC 的 RPMM	25	8	7.7	2.0
	50	5	5.6	1.32
DPBMM	25	5	5.16	0.93
	50	5	5.29	1.43

相应的 200 次 DPBMM 迭代的整体查准率 – 整体查全率曲线如图 5 所示.

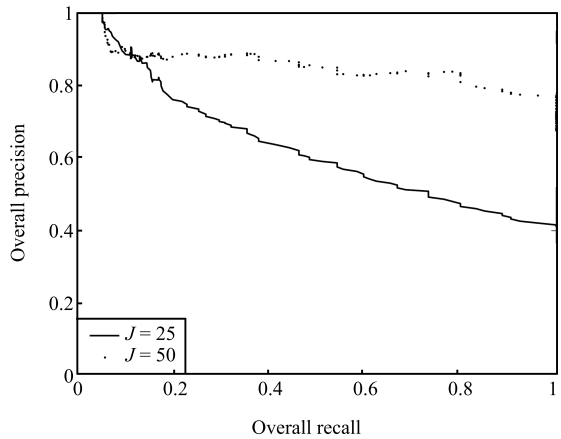

图 5 200 次 DPBMM 迭代的整体查准率 – 整体查全率曲线

Fig. 5 The curves of overall precision – overall recall in 200 DPBMM iterations

4 结论

本文提出一种用于处理 DNA 甲基化微阵列数据的聚类算法 DPBMM. 通过建立无限贝塔混合模型, 利用 “no-gaps” 算法解决了构建贝塔混合模型产生的似然函数和 Dirichlet 过程基分布为非共轭关系导致的积分不可求问题. DPBMM 算法借助 Dirichlet 过程中聚类混合模型个数先验取 ∞ 的特征, 挖掘出数据中隐藏的特征. 此方法能够有效地从数据和模型的角度自动找出聚类类别个数. 但考虑到类似 DNA 甲基化微阵列数据的高维特性, 将进一步深入考虑 DPBMM 与特征提取算法相结合的综合模型的研究. 考虑到改变贝塔分布的两个参数即可改变贝塔分布的形状, 该算法也可用于分布于 [0, 1] 之间的数据的聚类分析.

References

1 Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 1987, **196**(2): 261−282

2 Fan Shi-Cai, Zhang Xue-Gong. Progress of bioinformatics study in DNA methylation. *Progress of Biochemistry and Biophysics*, 2009, **36**(2): 143−150
(凡时财, 张学工. DNA 甲基化的生物信息学研究进展. 生物化学与生物物理进展, 2009, **36**(2): 143−150)

3 Jones P A, Baylin S B. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, 2002, **3**(6): 415−428

4 Ang P W, Li W Q, Soong R, Lacopetta B. BRAF mutation is associated with the CpG island methylator phenotype in colorectal cancer from young patients. *Cancer Letters*, 2009, **273**(2): 221−224

5 Siegmund K D, Laird P W, Laird-Offringa I A. A comparison of cluster analysis methods using DNA methylation data. *Bioinformatics*, 2004, **20**(12): 1896−1904

6 Houseman E A, Christensen B C, Yeh R F, Marsit C J, Karagas M R, Wrensch M, Nelson H H, Wiemels J, Zheng S C, Wiencke J K, Kelsey K T. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics*, 2008, **9**(1): 365

7 Zhang Lin, Liu Hui. A Clustering method based on Dirichlet process mixture model. *Journal of China University of Mining & Technology*, 2012, **41**(1): 159−163
(张林, 刘辉. Dirichlet 过程混合模型的聚类算法. 中国矿业大学学报, 2012, **41**(1): 159−163)

8 Zhou Jian-Ying, Wang Fei-Yue, Zeng Da-Jun. Hierarchical Dirichlet processes and their applications: a survey. *Acta Automatic Sinica*, 2011, **37**(4): 389−407
(周建英, 王飞跃, 曾大军. 分层 Dirichlet 过程及其应用综述. 自动化学报, 2011, **37**(4): 389−407)

9 Bouguila N, Ziou D. A Dirichlet process mixture of generalized Dirichlet distributions for proportional data modeling. *IEEE Transactions on Neural Networks*, 2010, **21**(1): 107−122

10 Kuan P F, Wang S J, Zhou X, Chu H T. A statistical framework for Illumina DNA methylation arrays. *Bioinformatics*, 2010, **26**(22): 2849−2855

11 Escobar M D, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 1995, **90**(430): 577−588

12 Pitman Jim. Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, 1996, **30**: 245−267

13 MacEachern S N, Müller P. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 1998, **7**(2): 223−238

14 Gelman A, Carlin J B, Stern H S, Rubin D B. *Bayesian Data Analysis* (Second edition). Boca Raton: CRC press, 2004

15 Amigó E, Gonzalo J, Artiles J, Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 2009, **12**(4): 461−486

张 林 中国矿业大学信息与电气工程学院讲师. 主要研究方向为生物信息处理和信号处理. 本文通信作者.
E-mail: laurenie.zhang@gmail.com
(**ZHANG Lin** Lecturer at the School of Information and Electrical Engineering, China University of Mining and Technology. Her research interest covers bioinformatics and signal processing. Corresponding author of this paper.)

刘 辉 中国矿业大学信息与电气工程学院讲师. 主要研究方向为生物信息处理技术和 miRNA 靶标预测. E-mail: lhcumt@hotmail.com
(**LIU Hui** Lecturer at the School of Information and Electrical Engineering, China University of Mining and Technology. His research interest covers bioinformatics and miRNA target prediction.)